Proceedings of the 1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020), pages 235–239
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Adapting a Welsh Terminology Tool to Develop a Cornish Dictionary

Delyth Prys
Language Technologies Unit, Bangor University
Bangor, Gwynedd, Wales
{d.prys}@bangor.ac.uk

Abstract

Cornish and Welsh are closely related Celtic languages and this paper provides a brief description of a recent project to publish an online bilingual English/Cornish dictionary, the Gerlyver Kernewek, based on similar work previously undertaken for Welsh. Both languages are endangered, Cornish critically so, but both can benefit from the use of language technology. Welsh has previous experience of using language technologies for language revitalization, and this is now being used to help the Cornish language create new tools and resources, including lexicographical ones, helping a dispersed team of language specialists and editors, many of them in a voluntary capacity, to work collaboratively online. Details are given of the Maes T dictionary writing and publication platform, originally developed for Welsh, and of some of the adaptations that had to be made to accommodate the specific needs of Cornish, including their use of Middle and Late varieties due to its development as a revived language.

Keywords: Cornish, Welsh, lexicography, language revitalization

1. Background

Welsh and Cornish are two closely related languages belonging to the P-Celtic language group. Welsh is a minority language with approximately 562,000 speakers in Wales (Office of National Statistics, 2012), but is the largest of the modern Celtic languages in terms of speaker numbers, with a vibrant youth culture and a strong movement for language revitalization. This is backed by government strategy (Welsh Government, 2017), and an action plan to provide language technologies for Welsh (Welsh Government 2018). Cornish was declared extinct at the end of the eighteenth century, but has been revived and kept alive since then by a small group of adult learners, with a recent upturn in interest in the language and some families now passing on the language to their children. Accurate numbers of speakers are difficult to ascertain, as there is no question on the number of Cornish speakers included in the ten yearly UK census. However, various estimates think that there may be in the region of 300 fluent speakers and 3,000 learners of the language with around 300,000 knowing a few words (Ferdinand, 2013; O'Neill, 2005).

In contrast to the extensive academic support Welsh enjoys, with long-established departments of Welsh language and literature at all Welsh universities, Cornish language scholarship depends to a large degree on the work of a dedicated group of researchers working mainly on a voluntary basis. They are grouped together as the Akademi Kernewek (Cornish Academy)[1], an independent body responsible for the development of the Cornish language and establishing standards for it. The Akademi Kernewek is organized into four panels, responsible for Dictionary Development, Signage and Place-names, Terminology, and Research, and works with Cornwall Council to deliver its agenda. It is recognized by Cornwall Council, the local authority in charge of the Cornish region, as the definitive body responsible for corpus planning for the Cornish language. Another positive development for the Cornish language has been the establishment of a Cornish language office at Cornwall Council, and a small grant from the UK government for Cornish language activities and resources.

Among the Cornish language community's core requirements was the creation of an up-to-date lexical resource, available in an online format, and easily expandable to include new content. Previously there had been various voluntary projects to create lexical resources, but lack of funding meant they were fragmented and difficult to coordinate and publish. There had also been four competing varieties of Cornish, with no generally accepted written standard. This was resolved in 2007 by the creation of a Standard Written Form (Bock and Bruch, 2008) which cleared the way for renewed action on resource creation.

A chance meeting between the Cornish Officer of Cornwall Council and researchers at the Language Technologies Unit, Bangor University, led to a collaboration to adapt an existing Welsh dictionary-writing and publication platform to the needs of the Cornish language and to port existing Cornish dictionary data to it, thus ensuring its speedy publication. It was also intended to aid its further development, future-proofing it so that a dispersed team of editors could continue working on the dictionary, in a way that was compatible with the needs of contributing scholars.

2. Technical Details

The dictionary-writing and publication platform developed at Bangor University was Maes T (Andrews & Prys, 2011). It was originally conceived in order to help terminologists and subject specialists who were geographically distant from each other to collaborate, using a friendly, easy-to-use interface, feeding into a secure and stable master database at its back-end. From this master database it would then be able to easily produce an online version, and if needed, to publish in other formats as well. It had been used to publish over twenty Welsh terminology dictionaries online, some with their own websites in addition to appearing together in a 'one-stop shop' on the Welsh National Terminology Portal (Prys, Jones and Prys, 2012). It had also been adapted to convert The Welsh Academy English-Welsh Dictionary (Griffiths and Jones, 1995), originally published on paper, to an electronic, on-line format (Welsh Language Board, 2012).

[1] More information on the Akademi may be found on their website https://www.akademikernewek.org.uk/

Both the terminology panels and the dictionary panels of the Akademi Kernewek share the Maes T system for work, and so a new fork was created from the original terminology orientated version of Maes T, rather than a further adaptation of the first fork created for Welsh Academy Dictionary digitization project. This Maes T interface guides the user through the input and decision making process, with different tabs leading from collecting candidate terms onwards to defining the meaning or concept, then deciding on the standardized form to finally inputting the linguistic information. To date, the Cornish dictionary and the terminology panels have been working on separate projects, but this is catered for within the Maes T system, where different dictionaries can be shown together in the interface for the published dictionary entries. This is similar to the way that the Welsh National Terminology Portal displays records from different dictionaries together on the same results page. Legacy data from work already done to prepare for the Gerlyver Kernewek were ported from other formats, after some initial pre-processing. To date there are 13577 Cornish entries in the main dictionary, with an additional 1400 entries in the terminology one.

Despite the close linguistic relationship between Cornish and Welsh, most of the new fields needed in the Maes T schema for Cornish belong to the Cornish linguistic information stage. The Cornish Dictionary, Gerlyver Kernewek (Akademi Kernewek, 2018) is written in the new Standard Written Form, but requires extra fields to show Middle and Late Cornish forms, as the revived language derives from sources from two different periods, roughly corresponding to the two broad time periods of Middle and Late. This also necessitates further adaptation to show the pronunciation of different forms if needed, as well as different plurals of some forms.

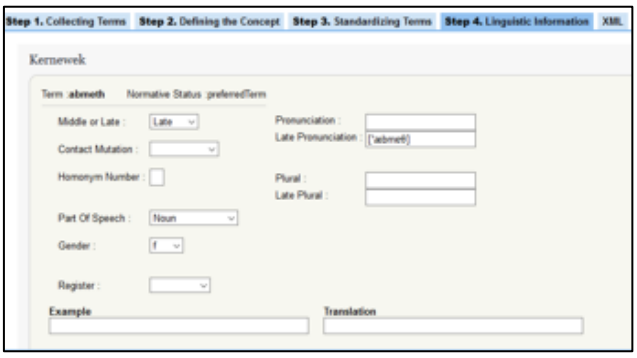

Figure 1: Screenshot of the Linguistic Information screen in Maes T showing fields for Middle or Late forms.

In addition to being able to cope with Middle and Late forms, another new field had to be introduced for a new part of speech, that of Collective Noun. Although this was present in earlier forms of Welsh it is no longer used as a part of speech category in the contemporary language, in contrast to Cornish. This category was therefore added, as well as fields for English glosses, etymology, attestations and example sentences, all requirements requested by our Cornish colleagues in order to capture the different types of information held in their legacy databases and documents.

3. Publishing Online

The publishing interface followed closely the interface already developed for the Welsh dictionaries. This allowed the input of words in either Welsh or English in the Welsh dictionaries, or Cornish or English in the case of the Gerlyver Kernewek, into a simple search box which would then show all relevant entries found. Language direction can be changed at the simple click of a button, and additional information on parts of speech is displayed through using a simple mouse over feature.

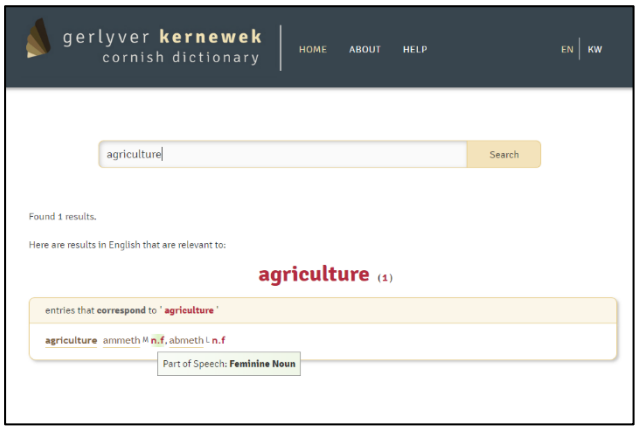

Figure 2: Screenshot of the Gerlyver Kernewek showing a simple search for ‘agriculture’ showing middle and late Cornish forms and the mouse over for part of speech.

In the Gerlyver, any underlined words can be clicked to show further information in a new view, enabling sophisticated searches with clear layouts.

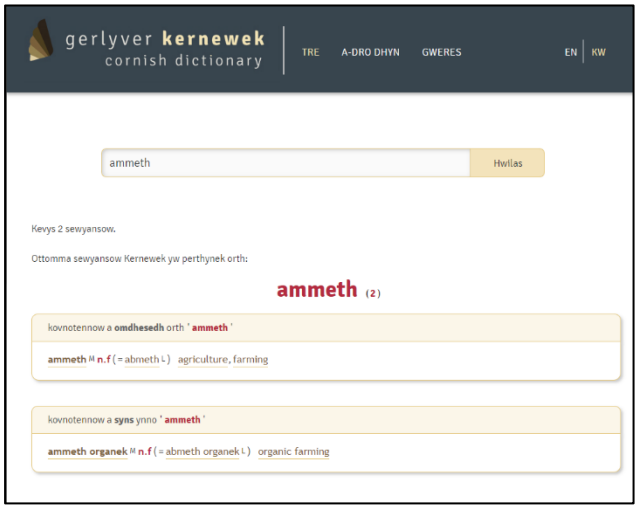

Figure 3: Screenshot of the Gerlyver Kernewek showing the result for ‘ammeth’ which can be accessed both by typing in ‘ammeth’ in the search box and by clicking on ‘ammeth’ underlined in the screenshot in Figure 2.

Electronic dictionary design has become important for many lexicography projects as dictionaries have moved increasingly to the digital sphere, with major languages such as English leading the way (Lew, 2010). However,

small language communities are still struggling to provide basic lexicographical resources for their language and may not have access to readily available technical expertise. The Digital Language Survival Kit names dictionary making as a vital, basic resource for language. Whilst acknowledging that building a dictionary is a task for experts, it accepts that this is not always possible, and names several open source platforms, including Wiktionary, as possibilities for collaborative efforts for language communities lacking the infrastructure to make large investments (Berger et al, 2018). Such international platforms are of necessity generic, and designed to be applicable to a broad range of languages.

Where investment or grant aid has been forthcoming, some communities have used it to employ commercial companies to advise or provide technical expertise. Although sustainability and upkeep of legacy data are important topics internationally (Wallnau et al, 2000; Almonaies et al, 2010), they are not always included in project plans for less-resourced languages, resulting in problems at a later date, as discussed in a Round Table on Celtic Language Technologies (Prys and Williams, 2019). Where commercial companies are employed, responsibility must rest with the commissioners of the original work to ensure its sustainability once the project ends. Some have argued that small language communities should insist on open licences for all their tools and resources, as articulated by openlt.org (n.d.) and their manifesto for open language technologies. However, even where free or open source software is used, there still needs to be planning for the long term, and a strategy for the future upkeep and development of the resource. Perhaps the most important element, whatever software is used, is that language communities are empowered to carry out lexicographical and resource development themselves, nurturing expertise within their community without needing high level technical knowledge, at least in the incipient stages. We therefore felt it important to design the Maes T interface to be user friendly to language experts who were unfamiliar with data input in a digital environment. We encouraged the Gerlyver team to take advantage of the more advanced features within Maes T where they were readily available and easy to use.

It is possible that a well-designed, easily navigable dictionary interface, as is true of other well-designed linguistic resources, can help raise the self-esteem of a language community that is used to being in a disadvantaged position in comparison to resources in the majority language. More research is needed to confirm this postulate, but some researchers, e.g. Tsunoda (2013) have argued that improved self-esteem in itself can help with language revitalization.

In the case of Maes T, there was already functionality to support the use of diagrams, scientific symbols, mathematical notations and photographic images within its definitions. These have proved useful to some Welsh terminology dictionaries developed in Maes T, but, to date, the main feature of interest for the Gerlyver Kernewek has been the inclusion of illustrations for some dictionary entries. This is particularly useful for names of plants and animals, and combined with the inclusion of the Latin scientific name, as used by the Akademi Kernewek terminology panel, whose entries also appear in the Gerlyver, have helped enrich them. In some instances the Welsh cognates have also been added, as the information could be shared with existing Welsh dictionaries. This is an on-going process, and other types of information may be added in future.

Although original images can be added by hand, we have found that appropriately licenced photographs from Wikidata Commons[2] are a useful resource of images, and can be imported into our dictionaries without too much trouble. They have been extensively used in the Welsh species dictionary Y Bywiadur (Llên Natur, n.d.), and are now used also in term entries in the Gerlyver. Public reaction since the online publication of the dictionary in June 2019 has been positive, with over 48,000 dictionary searches having been undertaken in it during the first two weeks. It has proved to be particularly popular with language learners both inside and outside of Cornwall, as a free online dictionary, due to the difficulties of distributing Cornish language resources to a wider audience.

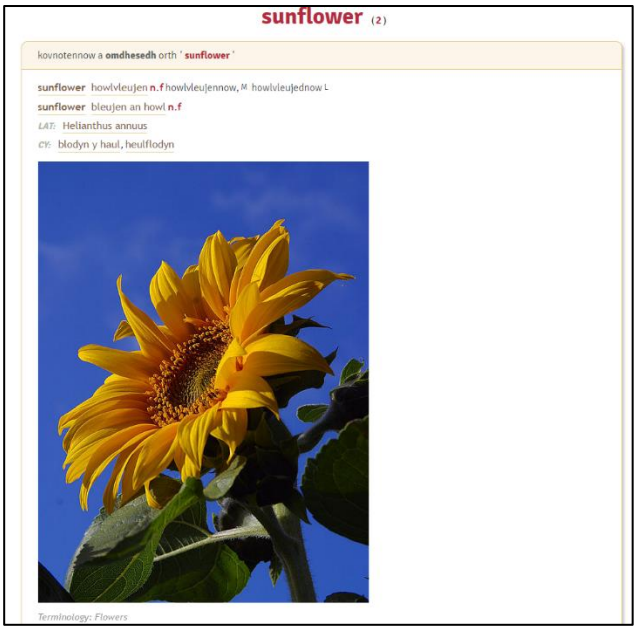

Figure 4: Screenshot showing part of the entry for 'sunflower' showing an illustration of the flower together with its Latin scientific name and Welsh cognates.

[2] Wikidata Commons is an online repository of free-use images, sounds and other media. https://commons.wikimedia.org/wiki/Commons:Wikidata

4. Further Work

Publishing a comprehensive online dictionary is only one element of the Cornish plan to provide modern language technology tools and resources for their language. The master database that lies behind the Maes T platform needs to be conceived of as a valuable resource in itself, not only as a way to produce an online dictionary. It can be used and reused to produce other lexicographical products, such as a phone or tablet dictionary app, similar to the Welsh Ap Geiriaduron[3], and an online word-by-word translator, similar to the Welsh Vocab (Jones, Prys and Prys, 2016). Other possibilities include using it to produce a Cornish spellchecker and language teaching aids. Teaching aids in particular are of particular interest to Cornish, given that most speakers are second language learners.

Sound files will shortly be added to the Gerlyver, to guide the pronunciation of Cornish. More ambitious plans include developing Cornish speech technology and machine translation, following recent and ongoing similar projects for Welsh (Jones and Cooper, 2016; Prys and Jones, 2019), and extending their corpus resources.

5. Conclusion

Many small language communities find it difficult to source both linguistic and technical expertise from within their own community. Collaboration can help overcome these difficulties, and the exchange of ideas benefits both parties. Building up expertise has wide-ranging benefits, foremost amongst which are a greater confidence and pride in our languages, which can help in revitalization efforts.

The collaborators were initially drawn together by a shared linguistic heritage, but found that other commonalities included a determination to develop sustainable resources and empower their own communities through the use of language technologies.

Building on the success of the Gerlyver Kernewek, we foresee a long-term partnership, where capacity will be built up within Cornwall, and a younger generation of both Cornish linguists and computational experts will be able to undertake their own resource and tool development, and where both Wales and Cornwall can contribute together to wider international projects.

6. Acknowledgements

We wish to thank Cornwall Council for funding the project and for their continued support and encouragement. Special thanks are due to Mark Trevethan, their Cornish Language Officer, and Davydh Trethewey who both provided invaluable help. We wish also to acknowledge the hard work of members of the Akademi Kernewek dictionary and terminology panels and their contribution to the revitalization of the Cornish language.

7. Bibliographical References

Akademi Kernewek. (2018). The Cornish Dictionary / Gerlever Kernewek. Akademi Kernewek, Cornwall. https://www.cornishdictionary.org.uk/ [accessed 13 February 2020].

Almonaies, A. A., Cordy, J. R., and Dean T. R. (2010) Legacy System Evolution towards Service-Oriented Architecture. School of Computing, Queens University Kingston, Ontario.

Andrews, T. and Prys, G. (2011). The Maes T System and its Use in the Welsh-Medium Higher Education Terminology Project. In T. Gornostay & J. Vasil (editors.) Proceedings of CHAT 2011: Creation, Harmonization and Application of Terminology Resources. Riga, Latvia, pp 49-50.

Berger, K.C, Hernaiz, A.G., Baroni, P., Hicks, D., Kruse, E., Quochi, V, Russo I., Salonen T., Sarhimaa, A, Soria, C. (2018) Digital Language Survival Kit: The DLDP Recommendations to Improve Digital Vitality. DLDP.

Bock, A. and Bruch, B. (2008). An Outline of the Standard Written Form of Cornish. Cornish Language Partnership, Cornwall. http://kernowek.net/Specification_Final_Version.pdf [accessed 13 February 2020].

Ferdinand, S. (2013). Brief History of the Cornish language, its Revival and its Current Situation. *E-Keltoi. Vol. 2, 2*. December, pp 199-227.

Griffiths, B. and Jones, D.G. (1995). Geiriadur yr Academi: The Welsh Academy English-Welsh Dictionary. University of Wales Press, Cardiff.

Jones, D.B. and Cooper, S. (2016). *Building Intelligent Digitial Assistants for Speakers of a Lesser-Resourced Language.* In Claudia Soria et al (editors). Proceedings of the LREC 2016 Workshop "CCURL 2016 – Towards an Alliance for Digital Language Diversity". Portorož, Slovenia, May. European Language Resource Association (ELRA), pp 74-79.

Jones, D.B., Prys, G. and Prys, D. (2016). *Vocab: a dictionary plugin for websites. In Teresa Lynn et al (editors).* Proceeding of the Second Celtic Language Technology Workshop. TALN 2016. Paris, pp 93-99.

Lew. R. (2010). Online dictionaries of English. In Fuertes-Olivera, Pedro A. and Henning Bergenholtz (eds), E-Lexicography: The Internet, Digital Initiatives and Lexicography. London/New York: Continuum, p 230–250.

Llên Natur (n.d.). Y Bwyiadur. Llên Natur, Cymdeithas Edward Llwyd, Cymru. https://www.llennatur.cymru/Y-Bywiadur. [accessed 14 February 2020].

Office of National Statistics. (2012). Language in England and Wales : 2011.

openlt.org (n.d). A Manifesto for Open Language Technology. https://openlt.org/ [accessed 23 March 2020].

O'Neill, D. (2005). Rebuilding the Celtic Languages : Reversing Language Shift in the Celtic Countries. Y Lolfa. Talybont, p242.

Prys, D., Jones, D.B. and Prys, G. (2012). The Welsh National Terminology Portal. Bangor University, Bangor. http://termau.cymru/ [accessed 13 February 2020].

Prys, D., and Williams, I. (2019). A roundtable discussion to promote a strategic vision for Celtic Language Technologies. Bangor University, Bangor, pp 4-5. http://techiaith.bangor.ac.uk/wp-content/uploads/2019/08/A-roundtable-discussion-to-

[3] The Ap Geiriaduron is available in iOS and Android versions from the App Store and Google Play.

promote-a-strategic-vision-for-Celtic-Language-Technologies.pdf [accessed 323 march 2020].
Prys, M. and Jones, D.B. (2019). Embedding English to Welsh MT in a Private Company. In Teresa Lynn et al (editors). Proceeding of the Celtic Language Technology Workshop. European Association of Machine Translation. Dublin, Ireland, pp 41-47.
Tsunoda, T. (2013) Language Endangerment and Language Revitalization. Mouton de Gruyter, Berlin, p168.
Wallnau, K., Seacord, R. C., and Robert, J. (2000) A Survey of Legacy System Modernization Approaches. Software Engineering Institute, Carnegie Mellon University, Pittsburgh.
Welsh Government. (2017) Cymraeg 2050 : Welsh Language Strategy. Welsh Government. Cardiff.
Welsh Government. (2018). Welsh Language Technology Action Plan. Welsh Government. Cardiff.
Welsh Language Board. (2012). Geiriadur yr Academi : The Welsh Academy English-Welsh Dictionary Online. https://geiriaduracademi.org/ [accessed 13 February 2020].